Homework for Lecture 14-15 Regression Analysis  
Sections 9.1 -9.4

Consider the following data set labeled Gala, which describe the number of species of tortoise on the various Galapagos Islands. There are 30 cases (Islands) and 7 variables in the dataset:

**Species:** The number of species of tortoise found on the island  

**Endemics:** The number of endemic species  

**Elevation:** The highest elevation of the island (m)  

**Nearest:** The distance from the nearest island (km)  

**Scruz:** The distance from Santa Cruz island (km)  

**Adjacent:** The area of the adjacent island (km$^2$)

The data were presented by Johnson and Raven (1973) and also appear in Weisberg (1985).
1. In the following analysis, we consider the linear relationship between **Elevation** and **Endemics**

![Scatterplot Matrix](image)

a) What is the sample correlation between these two variables: $r = 0.7929$? Would you say this a strong correlation?
   
   Since $r$ is quite close to 1, there is a quite strong positive correlation between the two variables.

b) From the JMP output report the 95% confidence interval for the population correlation between **Elevation** and **Endemics**. Is this correlation statistically significant? Why or why not?

   The 95% CI for the population correlation, $\rho$, is given by (0.6056, 0.8979). The correlation is statistically significant ($\rho \neq 0$), because 0 does **NOT** fall in the 95% CI.
c) According to the scatterplot, do the data appear to be linear? Which variable is the dependent variable and which is the independent?

The data appear to be linear, however, the variance/spread in the data seems to get larger as Elevation \((x)\) increases. Elevation is the independent variable while Endemics is the dependent variable.

d) State the meaning of the coefficient of determination in the context of the problem?

About 63% of the variation in Endemics can be explained by Elevation (according to this dataset).
Using the JMP output do a 6-step hypothesis test for the slope parameter, using a significance level of 0.05.

(a) $H_0$: $\beta_1 = 0$ vs. $H_1$: $\beta_1 \neq 0$ where $\beta_1$ is the population expected change in Endemics when Elevation increases by 1 unit.

(b) Level of significance $\alpha = 0.05$

(c) Test statistic: $t = \frac{b_1 - 0}{s_{b_1}} = \frac{0.05}{0.007}$ (sampling distribution under $H_0$ is $t$ with $30 - 2 = 28$ df)

(d) $t = 6.89$ and p-value $< 0.001$

(e) Reject $H_0$ because p-value $< 0.05$

(f) Conclude that there is a statistically significant linear relationship between Endemics and Elevation.

f) Provide the Regression equation for predicting Endemics from Elevation

Endemics = 7.18 + 0.05 Elevation

g) Using the regression equation, predict the Endemics value for the following Elevation: 250, 600, and 1200. What is the error associated with these predictions?

7.18 + 0.05 \cdot 250 = 19.98 \\
7.18 + 0.05 \cdot 600 = 37.18 \\
7.18 + 0.05 \cdot 1200 = 67.18

The associated error is given by the root mean square error: 16.95.
h) Do the residuals look healthy? Why or why not?

The residuals do not look healthy as they appear to fan out.

i) Are any assumptions of the model being appear to be violated? If so, which one(s)?

We can see that the variance in the $y$'s increases as $x$ increases. So the assumption that the variance is constant is violated.

2. Consider the following dataset known as stronx. These data were collected in an experiment to study the interaction of certain kinds of elementary particles on collision with proton targets. The experiment was designed to test certain theories about the nature of the strong interaction. The cross-section($\text{crossx}$) variable is believed to be linearly related to the inverse of the energy($\text{energy}$ - has already been inverted). At each level of the momentum, a very large number of observations were taken so that it was possible to accurately estimate the standard deviation of the response($\text{sd}$).

In the following analysis, we consider the linear relationship between $\text{crossx}$ and $\text{momentum}$.
a) What is the sample correlation between these two variables: \( r = -0.6477 \)?
Would you say this a strong correlation?
We could say that there is a slightly strong negative correlation.

b) From the JMP output report the 95% confidence interval for the population correlation between \textit{crossx} and \textit{momentum}. Is this correlation statistically significant? Why or why not?
The 95% CI for the population correlation, \( \rho \), is given by \((-0.9073, -0.0306)\). The correlation is statistically significant (\( \rho \neq 0 \)), because 0 does NOT fall in the 95% CI.

c) According to the scatterplot, do the data appear to be linear? Which variable is the dependent variable and which is the independent?
No, the data do not appear to be linear. Momentum is the \textit{independent} variable and \textit{crossx} is the \textit{dependent} variable.

d) State the meaning of the coefficient of determination in the context of the problem?
About 42% of the variation in \textit{crossx} can be explained by Momentum.
e) Using the JMP output do a 6-step hypothesis test for the slope parameter, using a significance level of 0.01.

(a) $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$ where $\beta_1$ is the population expected change in crossx when Momentum increases by 1 unit.

(b) Level of significance $\alpha = 0.01$

(c) Test statistic: $t = \frac{b_1 - 0}{s_{b_1}}$ (sampling distribution under $H_0$ is $t$ with 8 df)

(d) $t = -2.40$ and p-value = 0.0429

(e) Fail to reject $H_0$ because p-value $< 0.01$

(f) We conclude that there is not a statistically significant linear relationship between crossx and Momentum.

f) Provide the Regression equation for predicting crossx and momentum

$$\text{crossx} = 281.20 - 0.79 \times \text{momentum}$$

g) Using the regression equation, predict the crossx value for the following momentum values: 20, 75, and 120. What is the error associated with these predictions?

We cannot predict the values for crossx since the model is not statistically significant.

h) Do the residuals look healthy? Why or why not?

No, the data look unhealthy. Residuals show a clear pattern!

i) Are any assumptions of the model being appear to be violated? If so, which one(s)?

Data are not linear. In addition, the errors $(x_i, y_i)$ given $\beta_0, \beta_1$ do not look independent. It appears to be an association.